# Interpretable XGBoost Based Classification of 12-lead ECGs Applying Information Theory Measures From Neuroscience

Hardik Rajpal[1,*], Madalina Sas[1,*], Chris Lockwood[2], Rebecca Joakim[3], Nicholas S Peters[4], Max Falkenberg[1,4]

[1] Centre for Complexity Science, Imperial College London, London, United Kingdom
[2] Independent Researcher, Luxembourg
[3] Wexham Park Hospital, Frimley Park NHS Foundation Trust, Slough, United Kingdom
[4] ElectroCardioMaths Programme, Imperial College London, London, United Kingdom
[*] These authors contributed equally to this paper

## Abstract

*Automated ECG classification is a standard feature in many commercial 12-Lead ECG machines. As part of the Physionet/CinC Challenge 2020, our team, "Madhardmax", developed an XGBoost based classification method for the analysis of 12-Lead ECGs acquired from four different countries. Our aim is to develop an interpretable classifier that outputs diagnoses which can be traced to specific ECG features, while also testing the potential of information theoretic features for ECG diagnosis. These measures capture high-level interdependencies across ECG leads which are effective for discriminating conditions with multiple complex morphologies. On unseen test data, our algorithm achieved a challenge score of* $0.155$ *relative to a winning score of* $0.533$*, putting our submission in 24th position from 41 successful entries.*

## 1. Introduction

Among numerous diagnostic tools within cardiology, the 12-Lead ECG is arguably the most important. Unlike other cardiac imaging modalities, the 12-Lead ECG is unique in that it is often interpreted by doctors without a specialism in cardiology, who may not have extensive experience in ECG interpretation, and therefore might miss, or misdiagnose, specific ECG conditions, particularly those which are rare or have complex morphologies. For this reason, computer-aided ECG interpretation is an important clinical aid, having been built into commercially available ECG machines.

As part of the Physionet/Computing in Cardiology Challenge 2020, we were tasked with the automated classification of 12-Lead ECG signals. A full description of the challenge is at [1].

Given that automated ECG labeling is already commonplace in commercial ECG machines, our team, "Madhardmax", pursued two key motivations: (1) to develop an algorithm with interpretable predictions, and (2), to test a range of new and existing information theoretical (IT) features which have become popular in the neuroscience community.

## 2. Methods

Our approach uses a feature based classifier for each heart condition using the Python implementation of Extreme Gradient Boosting (XGBoost) [2].

### 2.1. Machine Learning Model

Gradient boosting is a supervised learning technique used to produce ensembles of decision trees incrementally by optimising a loss function. The current work uses XGBoost, a special case of gradient boosting. Decision trees bring the benefit of interpretability by means of decision analysis on the structure of the trees. Ensemble methods are scalable due to the diversity among constituent models, and can learn higher order interactions between features.

For each ECG condition, we trained an XGBoost ensemble using features extracted from the EGC signal to output the probability that the sample manifests that condition. The models were individually trained and evaluated on the combined datasets using cross-validation. A threshold was chosen for each class so that any probability exceeding the threshold will result in the sample being labeled with that condition.

#### 2.1.1. Class Imbalance

The provided training data features classes with positive samples ranging from hundreds to tens of thousands. Ma-

chine learning models built on such imbalanced datasets will generally yield predictions that are biased towards the most frequent classes.

To make best use of available learning data for rare conditions, we randomly undersample the majority class to match the population of the minority class for each binary classifier. This barters some accuracy for conditions with very few samples for more robust classification, and decreases the chances of overfitting.

## 2.2. Features

The raw ECG signal for each channel is filtered using a second order Butterworth band pass filter between 1.5 to 25 Hz. The filtered signal is then further de-noised using a "db4" discrete wavelet transform to remove high frequency noise components in the signal.

After preprocessing, features are extracted from each lead of the 12-lead ECG recording. Features come under four main categories: (1) standard ECG wave, interval and segment measurements, (2) power spectrum features, (3) IT measures, and (4) known patient metadata.

### 2.2.1. ECG waves, intervals and segments, and power spectra

For each lead, features are extracted corresponding to standard ECG waves, segments and intervals, e.g. QRS complex properties (height, width, skewness), QT and PR segment lengths, RR interval. For segment lengths and intervals, normalised values are computed by dividing by the average RR interval. Wave heights and prominances are normalised by diving by the absolute R peak height.

Normalised and raw power-spectra are also used as features, having shown promise in previous electrogram based classification tasks [3].

### 2.2.2. Lempel-Ziv Complexity

Lempel-Ziv Complexity (LZc) was first introduced as a measure of estimating the complexity of a finite sequence of numbers [4]. However, in recent years LZc has been successfully used in neuroscience to capture signal diversity of EEG/MEG neural time sequences [5–7]. LZc in these studies tends to correlate with subject awareness and varies significantly between states of conscious awareness.

For a given discrete signal $X$ of length $T$ with $d$ discrete symbols, LZc is estimated by sequentially scanning the signal and populating a dictionary of observed distinct patterns. The number of patterns in this dictionary denotes the complexity of the signal $C(X)$. The normalized LZc is given as,

$$LZ_{norm} = \frac{C(X)\log_d T}{T}. \quad (1)$$

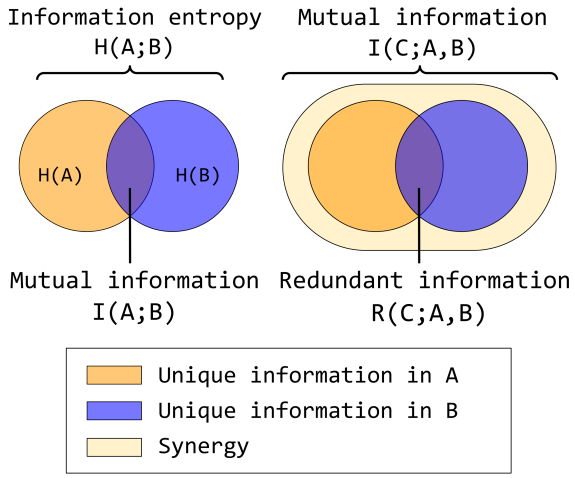


Figure 1. Comparison between bivariate and multivariate mutual information (MI). MI shared between all three variables corresponds to redundant information. The information that exists in a system of variables and is not unique to any one of them corresponds to the synergy of the system.

Inspired by the symbolic entropy analysis of ECG signals [8], we binarise the ECG signal based on the sign of the first order derivative of the signal. The binarised signal is then divided into non-overlapping windows of a specified length, and the average normalized LZc of each channel is estimated across these windows.

### 2.2.3. Synergy and Redundancy

Mutual Information (MI) is a well known IT measure of the amount of information shared between two signals [9]. Total Correlation (TC) and Dual Total Correlation (DTC) are the two multivariate extensions of MI that capture the higher order correlations that exist in a system of multiple correlated signals.

Recently [10], it was shown that these multivariate extensions can be used to quantify synergy and redundancy. An interaction between three or more variables is considered synergistic if there exists some unique information among the group of all variables which does not exist in the parts of the system, otherwise the interaction is deemed redundant, see Fig. 1. The synergy-redundancy of a given multivariate system is estimated using O-Information; the difference between TC and DTC [10]. Similarly, the sum of TC and DTC provides a measure of the strength of higher order correlations known as S-Information [10].

## 2.3. Optimisation and Evaluation

### 2.3.1. Cross Validation

In order to ensure the robustness of our classification algorithm, we employ K-fold cross validation. As the XGBoost algorithm uses a large number of parameters, we explored the best configuration by rotating through the K different possible splits and aiming for an increased F2 metric on the testing set.

### 2.3.2. Parameter Tuning

The performance of an ensemble increases training time as more models are added to the ensemble. The number of models in the ensemble (num_estimators), the maximum depth of the decision trees (max_depth), how many features are used for each tree, the learning rate (eta), the pruning (gamma) and regularisation (lambda) parameters can all affect the performance of the classifier. We used Grid Search to identify sets of promising parameters and used cross-validation to verify model robustness.

During training, we observed that some conditions require fewer features to classify, and that the tree and forest sizes vary greatly. Thus, we chose to use the more adaptable early stopping feature of XGBoost. The loss function minimised during the training phase is the Precision-Recall Area Under the Curve (AUPRC) [11], while the F2 metric was chosen for early stopping. Following optimisation, some parameters were fixed at:

{ max_depth: 10, eta: 0.1, gamma: 0, lambda: 1 }
Fixing any other parameters resulted in lower test scores.

### 2.3.3. Classification threshold optimization

In order to generate binary predictions from the XGBoost model output probabilities, two approaches were tested: (1) a subject is labelled with a condition if the prediction probability exceeds a fixed threshold of 0.9. This threshold is approximately optimal for achieving the maximum challenge score across the full training set. (2) A threshold is tuned for each condition.

To tune the binary prediction threshold, prediction probabilities for all 24 scored conditions are collated. First, the optimum threshold for the full set of conditions is found by performing a parameter sweep from 0 to 1. Next, a parameter sweep is performed for each individual condition, searching for an optimal threshold with respect to the scoring function, while retaining the global optimum for all other conditions.

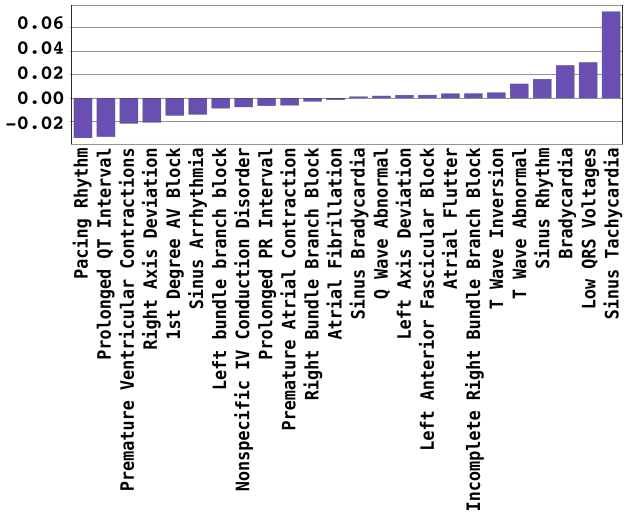


Figure 2. Taking the difference of feature entropy in two conditions $H(f_0) - H(f_0 + f_i)$ gives the amount of uncertainty reduced by adding the IT features ($f_i$) to the existing ECG features($f_0$). Conditions where we expect significant information redundancy between leads appear to show the greatest uncertainty reduction.

## 3. Results

### 3.1. Feature Importances

Cardiac conditions are usually diagnosed with very specific characteristics and ideally the binary classifiers should weigh those features appropriately and exclusively. Thus, a sparse feature importance distribution can be considered as a signature of precise feature selection by a model. Shannon Entropy ($H$) of any given distribution is a good measure of sparsity. In the normalized form it is given as,

$$H = \frac{-\sum_{i=1}^{i=N} p_i \log p_i}{\log N}, \tag{2}$$

where $N$ is the total number of states of the system. Shannon Entropy, a measure of uncertainty, is high when the probability mass is distributed across a wide range of states. It is maximum for the case when the probability of each state is equally likely. Conversely, it is minimum if one of the available states has probability 1. A feature importance distribution is obtained by normalizing the vector of feature gains, i.e. the average training loss reduction gained when using a feature for splitting. Shannon Entropy is then used to estimate the sparsity of the distribution.

### 3.2. Scores

The XGBoost ensemble was trained twice, once with IT measures, and once without. Table 1 shows standard

| | With IT | Without IT |
|---|---|---|
| AUPRC | 0.55 | 0.512 |
| Accuracy | 0.398 | 0.379 |
| Precision | 0.776 | 0.804 |
| Recall | 0.936 | 0.934 |
| F2 Score | 0.626 | 0.601 |

Table 1. Test metrics for the XGBoost models trained with and without IT measures on the training validation data. Binary predictions are derived using a threshold of $> 0.9$.

machine learning metrics calculated on the testing set in cross-validation. To improve the challenge metric, we have attempted to optimise the classification threshold.

On training data, threshold optimisation improved the challenge score with IT measures from 0.626 to 0.739 and from 0.601 to 0.731 without IT measures. For final scoring, the implementation with IT measures and threshold tuning was submitted resulting in a validation score of 0.533 and a final score of 0.155 on the full unseen test data. The significant score reduction from validation to final scoring suggests that significant over fitting hampered the final algorithm.

We cannot directly assess the impact of the IT measures on the final challenge metric. However, we note that in the F-scores provided by the challenge organisers, our algorithm ranks significantly higher for conditions with a large uncertainty reduction in the entropy, see Fig. 2, relative to our overall rank of 24th out of 41 teams (e.g. rank 2 for sinus rhythm).

## 4. Discussion and Conclusions

Our 12-Lead ECG classification algorithm aimed to achieve two goals: (1) to be interpretable - this is naturally the case using XGBoost decision trees and feature engineering. (2) To test the potential of information theoretic measures popularised in neuroscience as an effective diagnostic feature in ECG classification. Our results show that although IT measures are largely irrelevant for some conditions, there is notable uncertainty reduction for a number of conditions in which we expect global ECG commonalities across many different condition morphologies, indicating significant signal redundancy. Such conditions include conditions like sinus rhythm which are defined by a lack of irregular features across all twelve ECG leads.

Overall, our approach suggests that IT measures may be of interest for future ECG classification studies. However, our score in the current competition indicates that significant improvements are required to avoid overfitting and to achieve consistent results across diverse datasets.

Our code is available on Github[1].

[1] https://github.com/mearlboro/PhysioNet_12ECG


## Acknowledgments

M.F. acknowledges a Ph.D. studentship from the EPSRC, Grant No. EP/N509486/1. N.S.P. acknowledges funding from the British Heart Foundation (Grant Nos. RG/16/3/32175 and RE/18/4/34215) and the National Institute for Health Research Biomedical Research Centre. N.S.P. acknowledges funding from the Rosetrees Trust, Grant No. A1173/M577.


## References


[1] Perez Alday EA, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiol Meas 2020; (In Press).

[2] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; 785–794.

[3] McGillivray MF, et al. Machine learning methods for locating re-entrant drivers from electrograms in a model of atrial fibrillation. Roy Soc Open Sci 2018;5(4):172434.

[4] Lempel A, Ziv J. On the complexity of finite sequences. IEEE Transactions on Information Theory 1976;22(1):75–81.

[5] Schartner MM, et al. Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin. Sci Rep 2017;7(1).

[6] Mateos DM, et al. Measures of entropy and complexity in altered states of consciousness. Cogn Neurodyn 2017; 12(1):73–84.

[7] Dolan D, et al. The improvisational state of mind: A multidisciplinary study of an improvisatory approach to classical music repertoire performance. Front Psychol 2018;9:1341.

[8] Srinivasa M, Pandian P. Application of entropy techniques in analyzing heart rate variability using ECG signals. Int J Recent Innov Trends Comput Commun 2019;7:9–16.

[9] Cover TM, Thomas JA. Elements of Information Theory. Wiley Series in Telecommunications and Signal Processing. Wiley, 1991. ISBN 9780471062592.

[10] Rosas FE, et al. Quantifying high-order interdependencies via multivariate extensions of the mutual information. Phys Rev E Sep 2019;100:032305.

[11] Boyd K, et al. Area under the precision-recall curve: Point estimates and confidence intervals. In Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-40994-3, 2013; 451–466.



Address for correspondence:

Max Falkenberg
Centre for Complexity Science, Imperial College London,
London SW7 2AZ, United Kingdom
max.falkenberg13@imperial.ac.uk